THE PROPOSED USE OF REGISTRIES IN INFORMATION INTEGRATION
A White Paper of the
INFORMATION INTEGRATION NETWORK
Draft - June 27, 2000

(A more detailed version of this paper is available from the I-3 Team)

PURPOSE

This paper is designed to initiate a discussion that a set of linked registries can/should form a central element of EPA's information integration efforts. Registries - official and authoritative records of specific, well-defined data of interest to EPA - are needed to:

- Jump start an 'Enterprise Repository' - a network of trusted, reliable information that will serve as the core upon which programmatic information collections are connected and as a core record for multiple users. (These data will support many of the analyses upon which our business decisions will be based.)

- Offer the efficiencies and opportunities of shared data resources and eliminate or minimize the resource demands and inherent problems of managing separate, redundant data collections of like information.

- Provide the focus of EPA data quality assurance and control - a source of authoritative, reliable, trusted, secure information, replacing current data systems.

- Allow data users – including analysts in regional and program offices – to focus on use of data rather than on the data management concerns that many "stovepipe" systems impose on program systems. (Separating the rigors of managing data from managing its use in analyses will free up program resources and mean that modernization of data systems will not impede analyses.)

Registries Defined:

Registries, as used in this paper, are defined as **the official and authoritative records of specific, well-defined data of interest to EPA.** Registries will contain, for instance, definitive lists of official names, numbers, and/or other information used, or potentially used throughout the Agency as the authoritative record for that data. Registries will provide links to more detailed or less often used data housed outside the registry.

Data will be housed in registries - **managed by OEI and program offices -** to provide rapid and convenient access and will be the focus of data quality assurance. In addition, registries will be designed to be the focus of EPA data integration and enhance the Agency's commitment to a

consistent and non-duplicative core data holding and involve a shared commitment to data stewardship. Data known to be in registries would not be collected again, and would not be duplicated in other Agency (or perhaps Exchange Network) databases, but registries might be enlarged to include data that extends the utility of the basic data in the registry.

Envirofacts is a form of registry, one that is a composite of data from major program system databases, but also one that serves as the only repository of Toxic Release Inventory (TRI) and Risk Management Plans(RMPs).

Registries and Databases and Libraries:

Registries can be simple tables, groups of related tables, or databases that contain data that provide commonly used and shareable information on the subject for which the registry was established -- registries are common look up tables of shared information. In some cases they are more structured databases, but always serving as components of programmatic and corporate information collections. They are shared and corporately managed. Registries can be nestled within each other as components of a database management system as, for example, a Corporate Organization registry or an Industry Classification Registry that could logically be incorporated into the Facility Registry System. It is important that registries be housed in systems with compatible architectures (such as in Oracle formats accessible through SQL*Net). Registries will also provide links to more detailed data not included in the registry.

Registry Structure:

Each registry will be built around a data model that will describe the relationships between the data elements in the registry. A business plan for the registry will define who is responsible for developing, housing, and maintaining the registry.

Proposed Registries of Integrated Data:

- Regulated Facility Registry
- Site or Station Registry
- Substance Registry
- Organizations Registry
- Sector Registry
- Permit Registry
- Compliance Registry
- Regulation Registry
- Policy Registry
- Place Registry
- Studies Registry
- Ambient Monitoring Registry

The Use of Registries:

Registries will be archives.

Registries will structure the collection of new data.

Registries will make program support of analytical software easier.

Data Stewardship and Registries:

Users of registry information will be responsible for registry data quality. The Agency must ensure that the movement or replication of data into a registry is done with the cooperation of those who have traditionally been responsible for it - reinforcing the ethic of data stewardship and a commitment to an integrated EPA information infrastructure. (See also the White Paper addressing Shared Network Governance and Stewardship of Data and The Exchange of Data.)

Registries will initially be constructed from existing separate data systems using data models established for the purpose. Once the registries are established, the legacy systems of the EPA program offices (and particularly their revised newer versions) can be used as the analytic systems to interpret the data contained in the registries. Thus, registries begin to separate basic data from the software that helps convert that data into information. As registries are implemented, confusion over where the appropriate data resides can be controlled either through some form of synchronization between legacy data and registry data, or by absorbing the data system thereafter providing direct access to analytic programs to the registry. This separation of data from the analysis planned for it will facilitate EPA participation in the data exchange network and will also allow easier regional and program office analyses. This will allow wider access to what will become larger, more diverse, and more useful, EPA (and Exchange Network) data holdings. In the future, separate legacy analytic systems can be modernized, with registries as their sole or primary data source. Upgraded analysis programs may be less costly to build and maintain than those, as several are now, locked to the data they contain.

RECOMMENDATION/CONCLUSION

The QIC should encourage the exploration of using registries as an organizing element of EPA's participation in the Information Exchange Network.